título

Procesamiento masivo de datos mediante Cassandra y Spark

Máster en Sistemas Informáticos Avanzados
Septiembre de 2016

Autor:

Xabier Zabala Barandiaran

Supervisores:

German Rigau i Claramunt
UPV/EHU
Iñigo Etxabe y Beñat Aranburu
Datik Información Inteligente S.L.

Agradecimientos

En primer lugar, quisiera expresar mi gratitud hacia las personas que han posibilitado la concepción y el desarrollo de este proyecto. Agradezco a German Rigau i Claramunt, supervisor del proyecto por parte de la UPV/EHU, la predisposición mostrada y el asesoramiento ofrecido durante el transcurso del mismo. Doy también las gracias a Inigo Etxabe y Beñat Aranburu, supervisores del proyecto por parte de Datik Información Inteligente S.L., por poner a mi disposición todos los medios tecnológicos necesarios para llevar a cabo el proyecto y por el trato ofrecido desde el primer día.

Agradecerles, cómo no, a mis padres José Javier Zabala y María Pilar Barandiaran por el esfuerzo desempeñado para facilitarme, en la medida que les ha sido posible, el camino que he recorrido hasta llegar aquí. Me congratula haber sabido responder satisfactoriamente a la confianza que ellos siempre han depositado en mí. Por todo lo que han hecho y por todo lo que suponen para mí, un beso enorme para los dos.

No quisiera olvidarme de todas las personas que han estado a mi lado durante este maravilloso periplo, a los cuales no me atrevo a mencionar de forma individual por el miedo de dejar a alguno en el tintero. Gracias a los amigos de toda la vida por el apoyo ofrecido durante este camino. Gracias a los compañeros de la facultad, por todo los momentos vividos juntos y en especial a aquellos que durante esta etapa se han ganado a pulso el privilegio a ser parte importante de lo que me resta de existencia.

Por último, pero no por ello menos importante, quisiera evocar a todos los docentes que han tomado parte en mi formación desde aquel Septiembre del 2008 y agradecer a todos ellos el conocimiento compartido y el esfuerzo invertido en mí durante estos años.

Gracias de todo corazón a la gente mencionada en este breve capítulo por haber hecho de mí un mejor profesional y sobre todo una mejor persona.

Resumen

Proyecto Final del Máster en Sistemas Informáticos Avanzados. Estudio comparativo de carácter empírico realizado sobre el rendimiento ofrecido por varias tecnologías emergentes en el área del Big Data a la hora de operar en escenarios que requieren un almacenamiento y procesamiento eficaz de volúmenes masivos de datos.

Se ha construido un clúster compuesto por tres nodos virtuales que operan sobre Linux. Una vez configurados y dotados de tecnología necesaria para el correcto funcionamiento de MySQL Cluster [**mysqlcluster**] y Apache Cassandra [**apachecassandra**], se ha procedido a poblar dichos sistemas de almacenamiento utilizando un data-set público de aproximadamente 25GB. Un conjunto de consultas diseñadas atendiendo a la naturaleza de los datos almacenados han sido ejecutadas sobre ambas infraestructuras. El uso de Apache Spark [**apachespark**] durante el proceso ha posibilitado la distribución de la carga computacional entre los nodos que componen el clúster. A su vez, se ha erigido un nodo virtual con especificaciones técnicas equivalentes a la suma de los 3 nodos virtuales que forman el clúster para emular la respuesta ofrecida por una base de datos MySQL tradicional en el mismo contexto.

El estudio evidencia que, entre las tecnologías de almacenamiento distribuido comparadas, Apache Cassandra es la alternativa más eficaz a la hora de tratar volúmenes masivos de datos a costa de tener que realizar un análisis y diseño previo de los mismos. No obstante, MySQL Cluster es una opción muy a tener en cuenta, ya que sacrificando la eficacia de forma ligera, posibilita estructurar datos de manera más flexible. Ambas tecnologías demuestran estar mejor capacitadas que el MySQL monolítico tradicional para operar en el contexto analizado.

Palabras Clave: Big Data, MySQL Cluster, Apache Cassandra, Apache Spark, Clústering.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Desde Aristóteles y su libro Segundos analíticos hasta Galileo, padre de la ciencia moderna, pasando por científicos y filósofos como Leonardo da Vinci o Descartes, entre otros, han proclamado que un método de investigación basado en lo empírico y en la medición, sujeto a los principios específicos de las pruebas de razonamiento es el camino para alcanzar la verdad.

Los avances tecnológico de las últimas décadas unidos ha sido posible observar y medir ciertos fenómenos imposibles de realizar hasta la fecha La sociedad moderna vive inmersa en un trasiego de datos, cuyo volumen, lejos de estabilizarse, cada día va en aumento. El McKinsey Global Institute (MGI) estima que dicho volumen está creciendo un 40 % cada año y augura que entre 2009 y 2020 se verá multiplicado por 44[1].

Se trata de datos procedentes distintos ámbitos, creados por diferentes personas de todo el mundo, incluso por máquinas. Datos que es interesante almacenar y procesar para extraer una semántica de todo ese conjunto de caracteres, a priori inconexos entre sí.

(este fénomeno no ha pasado desapercibido en el ambito empresarial)Pasando de una escala global a una empresarial, la previsión sigue siendo muy pareja. Es por ello que, la gran mayoría de las empresas se hayan interesado en el Big Data. De un estudio realizado entre los altos ejecutivos de las firmas que lideran el Wall Street se desprende que el 96 % tiene planeadas ciertas iniciativas relacionadas con el Big Data, y el 80 % tiene finalizada alguna[2].

1.1. Contexto

Muchas veces nos encontramos con aulas con demasiados alumnos. Estas clases son especialmente comunes en los primeros cursos de los estudios universitarios, donde el número de alumnos alcanza fácilmente las tres cifras. Conocer todos los alumnos que tienen dudas y monitorizar a los alumnos para saber si todos o la mayoría han acabado el ejercicio son tareas difícilmente abordables por una sola persona. La mayoría de las veces se suelen ignorar los problemas y dudas, y se sigue adelante.

Los nuevos planes de estudio, que pretenden dejar atrás el sistema de educación mediante clases magistrales y dinamizar las clases, han supuesto un aumento en el número de clases prácticas y laboratorios que se realizan. Algunos centros han optado incluso por dividir las clases en grupos más pequeños para realizar las prácticas, pero muchas veces ésto no es posible. En esas situaciones el profesor acaba por no poder monitorizar completamente la clase.

Con el fin de tener un medio común se han implantado en los últimos años nuevas tecnologías en entornos docentes. Sin embargo, en muchos casos esta tecnología se limita a entornos de apoyo a la docencia más que al alumnado, siendo muy popular el sistema de gestión del aprendizaje Moodle. Además, el uso más frecuente de estos sistemas es el de simples almacenes de recursos bibliográficos (enlaces, apuntes, transparencias, etc.).

Por otro lado, la expansión de las tecnologías móviles y tabletas, con las que los alumnos están cada vez más familiarizados, no ha sido aprovechada. Estas tecnologías están ya mayoritariamente presentes en las aulas, la mayoría del alumnado dispone de alguno de estos dispositivos, pero su uso como herramienta educativa no es real, desperdiciando así todo su potencial como sistema de ayuda al aprendizaje. Es más, muchas veces el uso de estos dispositivos está prohibido o limitado en clase.

1.2. Propuesta

Nuestra propuesta pretende modificar y actualizar los modelos educativos presenciales a través de herramientas que faciliten la captura de la información de lo que sucede en el aula, con el objetivo de proporcionar *feedback* a profesores y estudiantes sobre su progreso en el aprendizaje. Esta propuesta se materializa en la aplicación *PresenceClick* que facilita la captura colaborativa de esta información entre alumnos y profesores de manera ágil.

PresenceClick es un entorno modular colaborativo que facilita el registro en tiempo real

de la información sobre los alumnos en sesiones tradicionales de aprendizaje con el propósito de obtener feedback de ayuda para mejorar el proceso de enseñanza aprendizaje.

PresenceClick está compuesto por un sistema web (*webClick*), una aplicación móvil (pClick) y un servidor de datos para mantener el modelo del Alumno y el modelo de Grupo (*Learner and Group Models*). Ambas plataformas son a su vez modulares, dividiéndose así en diversos módulos encargados de las distintas interacciones que se registran en clase. La plataforma *pClick* captura las interacciones de aprendizaje en tiempo real. Por otra parte, *webClick* controla y muestra la información capturada durante sesiones cara a cara y permite a profesores y estudiantes seguir el progreso de sus alumnos durante el curso mediante visualizaciones y sus características *web responsive* permiten su acceso desde cualquier dispositivo, por ejemplo: PC, portátiles, tabletas o *smartphones*. La información capturada e integrada en los modelos de Alumno y de Grupo permite a profesores y estudiantes visualizar las actividades, conocimiento y comportamiento de estos. Este entorno cubre principalmente dos objetivos: mostrar a los profesores el estado de aprendizaje de sus alumnos y otras características para ayudarlos a adaptar sus estrategias de enseñanza; y mostrar a los alumnos su propio progreso y su comparación al grupo para promover un estado de reflexión que les permita mejorar su aprendizaje.

Actualmente, el Modelo Interacciones de *PresenceClick* incluye varios tipos de interacciones: de gestión (asistencia, evaluaciones), de actividad (preguntas de profesores y respuestas de alumnos), emocionales (sensaciones de los alumnos mientras realizan actividades de aprendizaje) y de comportamiento (participación, actividad).

La figura **??** muestra las líneas de interacción en la arquitectura de *PresenceClick*: (1) los asistentes comunican su presencia en clase al modelo del alumno a través de dispositivos de control de presencia, (2) profesores y estudiantes capturas las interacciones que realizan mediante *pClick*, y (3) los profesores y estudiantes interactúan con *webClick* para obtener información sobre el progreso del estudiante y del grupo.

Actualmente pClick cuenta con un módulo llamado *qClick* [**qclick**]: un sistema de Pregunta-Respuesta en el aula, en el que el profesor lanza una pregunta a los alumnos y ellos a través de sus móviles pueden contestar a esa pregunta. Los resultados visualizados de forma gráfica a las preguntas realizadas pueden mostrarse a los alumnos en tiempo real. Este sistema permite motivar el debate en clase, bien antes de lanzar la pregunta para que discutan entre ellos las posibles respuestas o bien después de responderlas observando los resultados de la clase.

En este proyecto, nuestro objetivo es crear **un nuevo módulo de *pClick* para capturar las interacciones entre profesor-alumnos en sesiones de ejercicios**. Por el momento

se desarrollará como una aplicación independiente y más tarde se integrará en *pClick*.

En esta aplicación el profesor dispondrá de una interfaz a la que accederá mediante su dispositivo móvil o tableta en clase e indicará a sus alumnos los ejercicios a realizar (está actividad se podrá planificar previa a la clase). Por su parte, los alumnos (que tienen que haber fichado con su tarjeta de alumno para entrar en clase) con sus dispositivos móviles (smartphones o tabletas) recibirán las notificaciones de los ejercicios a realizar y podrán indicar para cada uno si tienen dudas en su realización o si lo han terminado. El profesor podrá disponer en tiempo real de información sobre el porcentaje de alumnos que lo han realizado, alumnos que indican problemas en su realización y aquellos alumnos que no indican nada. Además el profesor podrá acercarse a comprobar y revisar las soluciones de los alumnos que indican haber terminado el ejercicio, y valorar su nivel de corrección o satisfacción en la realización, añadiendo las notas oportunas en el sistema que le permitirá seguir la evolución de cada uno de sus alumnos durante el curso. También podrá acercarse a aquellos que señalan problemas en su realización, con el fin de ayudarlos y evitar dificultades en su progreso.

Bajo este contexto surge **exerClick**, la herramienta para seguimiento de ejercicios en el aula. Esta herramienta, con todas sus funcionalidades, nace con el propósito de tener una visión más real de lo que hacen los alumnos, tanto una visión global del grupo como una individual y está dirigida a profesores y a los propios alumnos. De esta manera, el docente puede ofrecer un aprendizaje más adaptado e individualizado, aunque los grupos de alumnos sean muy grandes.

1.3. Organización del documento

En esta memoria se ha documentado el desarrollo de la herramienta ***exerClick***, dentro del Trabajo de Fin de Grado (TFG) del autor. En el documento se describe la propuesta, la planificación y gestión que esta lleva consigo, la implementación llevada a cabo y las conclusiones finales.

En este primer capítulo se ha introducido el problema a resolver y se ha explicado la propuesta presentada en este proyecto.

En el capítulo 2 se presenta el Documento de Objetivos de Proyecto (DOP). Este recoge el alcance y las fases y tareas del proyecto, el análisis de riesgos y el análisis de factibilidad.

Una vez en el capítulo 3 se explica la gestión llevada a cabo durante el proyecto. Se

presentan las metodologías utilizadas: Metodologías Ágiles e InterMod (adaptada a las necesidades de este proyecto). A continuación se detallan cada una de las iteraciones llevadas a cabo (como parte de la metodología InterMod): duración, objetivos y tareas realizadas. Al final del capítulo se muestra la documentación asociada a las iteraciones y los objetivos, además del seguimiento de tiempo realizado.

A continuación, en el capítulo 4 se detalla el análisis de requisitos. Primero se detallan los requisitos no-funcionales y luego los funcionales (prototipos en papel llevados a cabo durante las primeras iteraciones que dan una visión global del proyecto).

En el capítulo 5 se explica el diseño e implementación llevados a cabo. Se comienza mostrando la estructura de documentos del proyecto, luego el diseño realizado en base al análisis de requisitos del capítulo 4 y finalmente una visión general de la implementación de la lógica de negocio.

Para finalizar, en el capítulo 6 se presentan las conclusiones, líneas futuras para el proyecto y las lecciones aprendidas.

Fuera de la estructura general de la memoria, tenemos la bibliografia y los apéndices. En estos últimos tenemos las actas de reuniones, las actas de pruebas y la vista de relaciones de la base de datos (de la parte utilizada o creada específicamente para el proyecto).